Electronic Version
Stylesheet Version v1.1.1

Claims

[c1] 1. A method for identifying output documents similar to an input document, comprising:

(a) identifying a predefined number of keywords from a first list of rated keywords extracted from the input document to define a list of best keywords; the list of best keywords having a rating greater than other keywords in the first list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency;

(b) formulating a query using the list of best keywords;

(c) performing the query to assemble a first set of output documents;

(d) identifying lists of keywords for each output document in the first set of documents;

(e) computing a measure of similarity between the input document and each output document in the first set of documents;

(f) defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value;

wherein the list of best keywords has a maximum num-

ber of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency.

[c2] 2. The method according to claim 1, wherein each document in the second set of documents is identified as being one of a match, a revision, and a relation of the input document.

[c3] 3. The method according to claim 2, further comprising (g) if the second set of document contains an insufficient number of output documents, performing query reduction by removing at least one keyword in the list of best keywords that is not the keyword that is identified as belonging to a domain specific dictionary and having no measurable linguistic frequency.

[c4] 4. The method according to claim 3, further comprising if after performing (g) the second set of document contains an insufficient number of output documents, performing (h):
replacing the list of best keywords using keywords having a rating greater than other keywords in the first list of rated keywords; and
repeating (b) – (f).

[c5] 5. The method according to claim 4, further comprising (i) if the second set of documents includes a matching document but no similar documents repeating (a) - (g) using the matching document to identify similar documents.

[c6] 6. The method according to claim 5, performing (i) when textual content in the input document is identified using OCR or a portion of the input document matches the output document.

[c7] 7. The method according to claim 5, wherein the predefined number of keywords identified from the first list of rated keywords is five.

[c8] 8. The method according to claim 1, further comprising:
receiving an input document having textual content and image content;
performing OCR on the image content to identify text;
analyzing the text and the textual content to identify keywords.

[c9] 9. The method according to claim 1, further comprising:
recording a digital image representation of the input document;
performing OCR on the digital image representation to identify text;

analyzing the text to identify keywords.

[c10] 10. The method according to claim 1, further comprising:

(g) extracting from the input document the first list of keywords;

(h) determining if each keyword in the first list of keywords exists in a domain specific dictionary of words;

(i) for each keyword in the first list of keywords, determining its frequency of occurrence in the input document, also referred to as its term frequency;

(j) for each keyword identified at (h) that exists in the domain specific dictionary of words, assigning each keyword its linguistic frequency if one exists from a database of linguistic frequencies defined using a collection of documents, and assigning its linguistic frequency to a predefined small value if one does not exist in the database of linguistic frequencies;

(k) for each keyword that was not identified in the domain specific dictionary of words at (h), assigning each keyword its linguistic frequency if one exists in the database of linguistic frequencies; and

(l) for each keyword in the first list of keywords to which a term frequency and a linguistic frequency are assigned, computing a rating corresponding to its importance in the input document that is a function of its frequency of

occurrence in the input document and its frequency of occurrence in the collection of documents.

[c11] 11. The method according to claim 10, for each keyword that was not identified in the domain specific dictionary of words at (h) and that was not assigned at (j) a linguistic frequency from the database of linguistic frequencies, assigning each that matches a regular expression from a set of regular expressions a predefined rating.

[c12] 12. The method according to claim 11, further comprising, for each keyword in the first list of keywords, modifying the term frequency of keywords determined at (i) to a predefined maximum.

[c13] 13. The method according to claim 12, wherein keywords include phrases of keywords.

[c14] 14. The method according to claim 11, wherein the rating is a weight computed using the following equation:

$W_{t,d} = F_{t,d} * \log (N/F_t)$, where:

$W_{t,d}$: the weight of term t in document d;

$F_{t,d}$: the frequency occurrence of term t in document d;

N : the number of documents in the collection of documents;

F_t : the document linguistic frequency of term t in the collection of documents.

[c15] 15. The method according to claim 11, wherein keywords that do not match a regular expression from the set of regular expressions are removed from the first list of keywords.

[c16] 16. A method for computing ratings of keywords extracted from an input document, comprising:
(a) determining if each keyword in the list of keywords exists in a domain specific dictionary of words;
(b) determining a frequency of occurrence in the input document for each keyword in the list of keywords, also referred to as its term frequency;
(c) for each keyword identified at (a) that exists in the domain specific dictionary of words, assigning each keyword its linguistic frequency if one exists from a database of linguistic frequencies defined using a collection of documents, and assigning its linguistic frequency to a predefined small value if one does not exist in the database of linguistic frequencies;
(d) for each keyword that was not identified in the domain specific dictionary of words at (a), assigning each keyword its linguistic frequency if one exists in the database of linguistic frequencies; and
(e) for each keyword in the list of keywords to which a term frequency and a linguistic frequency are assigned, computing a rating corresponding to its importance in

the input document that is a function of its frequency of occurrence in the input document and its frequency of occurrence in the collection of documents.

[c17] 17. The method according to claim 16, wherein the keywords in the list of keywords are used to carry out one of language identification, indexing, categorization, clustering, searching, translating, storing, duplicate detection, and filtering.

[c18] 18. A system for identifying output documents similar to an input document, comprising:
a memory for storing the output documents and the input document and processing instructions of the system; and
a processor coupled to the memory for executing the processing instructions of the system; the processor in executing the processing instructions:
(a) identifying a predefined number of keywords from a first list of rated keywords extracted from the input document to define a list of best keywords; the list of best keywords having a rating greater than other keywords in the first list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency;
(b) formulating a query using the list of best keywords;
(c) performing the query to assemble a first set of output

documents;

(d) identifying lists of keywords for each output document in the first set of documents;

(e) computing a measure of similarity between the input document and each output document in the first set of documents;

(f) defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value;

wherein the list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency.

[c19] 19. The system according to claim 18, wherein the processor in executing the processing instructions further comprises:

(g) extracting from the input document the first list of keywords;

(h) determining if each keyword in the first list of keywords exists in a domain specific dictionary of words;

(i) for each keyword in the first list of keywords, means for determining its frequency of occurrence in the input document, also referred to as its term frequency;

(j) for each keyword identified at (h) that exists in the domain specific dictionary of words, means for assigning each keyword its linguistic frequency if one exists from a database of linguistic frequencies defined using a collection of documents, and assigning its linguistic frequency to a predefined small value if one does not exist in the database of linguistic frequencies;

(k) for each keyword that was not identified in the domain specific dictionary of words at (h), means for assigning each keyword its linguistic frequency if one exists in the database of linguistic frequencies; and

(l) for each keyword in the first list of keywords to which a term frequency and a linguistic frequency are assigned, means for computing a rating corresponding to its importance in the input document that is a function of its frequency of occurrence in the input document and its frequency of occurrence in the collection of documents.

[c20] 20. An article of manufacture for identifying output documents similar to an input document, the article of manufacture comprising computer usable media including computer readable instructions embedded therein that causes a computer to perform a method, wherein the method comprises:

(a) identifying a predefined number of keywords from a first list of rated keywords extracted from the input doc-

ument to define a list of best keywords; the list of best keywords having a rating greater than other keywords in the first list of keywords except for keywords belonging to a domain specific dictionary of words and having no measurable linguistic frequency;

(b) formulating a query using the list of best keywords;

(c) performing the query to assemble a first set of output documents;

(d) identifying lists of keywords for each output document in the first set of documents;

(e) computing a measure of similarity between the input document and each output document in the first set of documents;

(f) defining a second set of documents with each document in the first set of documents for which its computed measure of similarity with the input document is greater than a predetermined threshold value;

wherein the list of best keywords has a maximum number of keywords less than the number of keywords in the list of best keywords that are identified as belonging to a domain specific dictionary of words and having no measurable linguistic frequency.